Testing task difficulty evaluating parameters and identifying gestures as a valid indicator.

Renia Lopez-Ozieblo

Abstract

In second language acquisition, tasks based on various types of inputs are very popular. These inputs can be textual, aural or visual (or a combination of all three). Perceptions of task difficulty varies from student to student and assessing the complexity of the task can be a challenge to designers. This study investigates how ten Hong Kong participants, second language speakers of English, ranked the difficulty of three tasks based on different input modalities –textual, aural and visual. It also compares participants' rankings to those calculated through a number of parameters used to evaluate speakers' speech performance (Skehan, 2009; Robinson, 2011). In addition, this study explores the validity of new parameters based on gestures.

A meta-analysis of gesture studies confirms that gestures, movements of the hands and arms when speaking, are used by speakers to reduce cognitive load but also to help listeners understand the message (Hostetter, 2011). Different modality inputs might be imposing different cognitive loads, both in the processing of the information and in its transformation into speech/gesture. If there is a link between cognitive load and gesture, through studying gestures we might be able to find out more about the cognitive loads imposed by different modalities. Our hypothesis that there would be more gestures in the narrations of the more demanding tasks was confirmed by this study. This suggest that a gesture based parameter might be a good indicators of task difficulty.

Keywords: gestures; multimodal tasks; cognitive load; second language acquisition (SLA), evaluating parameters

1. Introduction

Task based language-learning developed during the late 80s (Candlin, 1987) as a more authentic and interactive approach to second language acquisition (SLA), teaching not just vocabulary and grammar but how to express and interpret meaning (Norris, 2016). Learners have different learning styles, so tasks tend to combine input modalities, written text, aural and visual. The benefits of the various modalities, how their combination aids learning and how to identify the difficulties they

present to the learner is an area still under study (Chen & Wu, 2015). Learners' perceptions of task difficulty seem to be related to how well they perform (Robinson 2011), with general patterns indicating that lower performances are related to higher cognitive-demand tasks (Norris, Brown, Hudson & Bonk, 2002). However, there is a misalignment of task designers and learners' perceptions of task difficulty (Elder, Iwashita, McNamara, 2002) and despite a number of studies manipulating various task factors, the difficulties imposed by the cognitive demands of language tasks are still debated (Lambert, Kormos & Minn, 2017; Révész Révész, Michel & Gilabert, 2016; Robinson, 2011; Sasayama, 2016; Skehan, 2009). In this study, we explored the cognitive demands imposed on second language speakers by different modalities of input. The tasks are three narrations based on input in three different modalities: a written text (textual), an audio (aural) and a video (visual). We investigated the validity of various well established measures used to evaluate the performance of language students, and thus the difficulties imposed by the tasks, and two new ones based on gestures.

After decades of discussion as to the function of gestures, whether communicative (for the interlocutor) or cognitive (for the speaker), there is now enough evidence of gestures aiding both functions. In a meta-study of gestures, Hostetter (2011) concluded that iconic gestures (of a representational nature) aid the interlocutor with the comprehension of spatial and motor ideas, whether scripted or spontaneous. On the other hand, production of gestures also helps speakers in tasks of comprehension, memorization and learning. Gestures co-occurring with speech are thought to lighten the cognitive load (Goldin-Meadow 2010), in particular if the content, or the input, is of a spatial nature (Kita & Davies, 2009). So far in language studies, gestures have not been integrated with other linguistic measures to determine task complexity (studies tend to focus on gestures alone). Therefore, we sought to identify whether they could be valid indicators of task difficulty.

2. Cognitive load

Cognitive load is a theoretical concept that refers to the demands on working memory to select, organize and integrate new information into an existing knowledge base (Mayer & Moreno, 2003). Cognitive load theory (Chandler & Sweller, 1991) is a framework that relates to working memory and its relationship with long term memory in learning and problem-solving (Diao, Chandler & Sweller, 2007).

Working memory theories are based on Baddeley and Hitch's model (1974). This states that immediate, working memory is a set of systems that includes two independent processors for words (both written and verbal) and visual information. These have coding, storing and retrieval capabilities that are orchestrated by a central executive function (see Baddeley, 1999 for a thorough explanation). The two verbal and visual systems are known as the phonological loop and the visuo-spatial sketch.

Learning styles are not a limitation as to what can be processed from a specific modality (Miller, 2001). Preference for one learning style or another means the recoding of the input into the preferred modality. A visualizer will just transfer the verbal content into imagery, and a verbalizer the images into words for further processing. Mayer and Massa (2003) point out that the various categories of learning styles often refer to different elements: cognitive ability (low or high spatial abilities to do something); cognitive style (whether thinking in images or words, related to the processing and representation of information); and learning preference (whether the learner prefers the data to be presented as text or graphics). Although a number of studies report enhanced students' performances when the modality of input matches their learning style (Ford & Chen, 2000; Surjono, 2015), Mayer and Massa (2003) point out that it is still not clear how these various elements relate to multimedia learning.

2.1. Assessment of cognitive load

The cognitive load cannot be directly observed and needs to be assessed in terms of its components, mental load, effort and performance (Diao, Chandler & Sweller, 2007). The most accurate measure of cognitive load is given as a combination of mental effort, which refers to the cognitive processing involved in the task, and performance which can be measured by additional post-tasks that test right answers. Mental load relates to the difficulty of the context and the task.

The complexity of the task is not a strict measure of the cognitive load imposed, as there are idiosyncratic factors that affect cognition efforts. The cognitive demands of a task go beyond its modality and complexity, needing to take into account the individual learner's skill and experience (Révész, et al., 2016). Methods of assessing cognitive load can be direct and indirect, subjective or objective. Objective direct methods include eye tracking, fMRI and dual-tasks (where participants have to remember a string of numbers or a similar memory task). Indirect methods measure the effort the task requires from the participant via changes in physical indicators, such as cardiovascular changes, pupil dilation, or electrical activity in the brain. Subjective methods

include those where the participant self-reports levels of stress or effort, often using Likert-type scales, or provides verbal reports. Subjective evaluation methods are non-intrusive and as they have been found to be valid they are widely used to measure task effort (Kollöffel, 2012; Kruger, Hefer & Matthew, 2014; Nesbit & Hadwin, 2006; Paas, Tuovinnen, Tabbers & Van Gerven, 2003; Révész et al., 2016; van Gog, Paas, van Merrienboer & Witte, 2005).

In second language acquisition, two hypotheses have been proposed to identify the factors related to the complexity of tasks and thus their potential difficulty to the learner, Skehan's (2001, 2009) trade-off and Robinson's (2011) cognition hypothesis. Both, although differing in the processes contributing to complexity, predict an impact on the quality of the language produced with either more demanding (Skehan, 2001) or more complex tasks (Robinson, 2011). Skehan (2009) bases his speech processing modal on that of Levelt's (1993) where speech is first conceptualised at the Conceptualizer module, then formulated at the Formulator and finally articulated. Furthermore, Skehan (2009) contends that each element –complexity, accuracy, fluency– will be developed at the expense of another, placing more or less strain on the Conceptualizer or the Formulator modules. The manipulation of information will tax the Conceptualizer; rarer lexical items will result in more errors, taxing the Formulator; however, clear macrostructure will result in fluency and accuracy as the Conceptualizer is hardly used.

It is understood that although task complexity does not equate with cognitive load, the difficulty experienced by the student in completing the task does. The complexity, accuracy and fluency of the language produced during these tasks is analysed to obtain an indication as to the difficulty imposed by them (Skehan, 2009; Skehan and Foster, 1997). This can be achieved by measures such as the number of pauses, disfluencies and repairs, measuring correction; the speed of delivery, measuring fluency; and the number of clauses, words in each clause and variety in the vocabulary used (token type ratio) measuring complexity (Chen, Ruiz, Choi, Epps, Khawaja, Taib, Yin, & Wang, 2012; Khawaja, Chen & Marcus, 2010; Gilabert, Barón, & Levkina, 2011; Robinson, 2011; Skehan, 2009).

Gestures might be used as an additional measuring parameter. The heavier the cognitive load imposed by a task (intrinsic load) or the modality of the input (extraneous load) the harder the processing of the data would be. This in turn would mean the involvement of more working memory elements (germane load) such as visuo-spatial and verbal. Cognitive load might be imposed by the challenges found in understanding the events, at the Conceptualizer level, or in

processing input into output, at the Formulator level. Although the processes leading to gesture production in these two cases might be different (conceptualization and formulation) the overall result is expected to be similar, a higher gesture rate.

3. Gestures

Gestures –for the purposes of this study: intentional movements of the hand and arms co-occurring with speech– are a type of kinesthetic movement that has been found to facilitate learning (Goldin-Meadow, 2003), as well as to aid recall in enacted sentences (Engelkamp & Cohen, 1991; Engelkamp & Zimmer, 1985; Toumpaniari, Loyens, Mavilidi, Paas, 2015).

Most gesture scholars defend a thought-gesture-speech-link. Either the thought is developed as the gesture and the speech are generated (Lopez-Ozieblo & McNeill, 2017; McNeill, 2015), or speech and gesture are processed independently, but combined at the thought level (Kita, 2000). Trofatter, Kontra, Beilock and Goldin-Meadow (2015) suggest that gestures generate mental images –rather than reflect them– grounded on the embodiment of physical properties (movement or aspect), and it is these that influence the thought. An alternative theory advances that gestures are the result of a process unrelated to that of speech and are mostly used to aid lexical retrieval (Rauscher, Krauss & Chen, 1996). In cases where speakers are experiencing lexical retrieval issues more gestures have been observed (Beattie & Shovelton, 1999; Frick-Horbury and Guttentag, 1998), however, these gestures are both for the benefit of the speaker and for the interlocutor, as they also indicate that the floor is still taken.

One of the earlier defenders of the gesture-thought link and the cognitive role of gestures was also Baddeley, who together with Hitch developed the working memory model mentioned above (1974). Baddeley proposed that gestures activated mental images in working memory (1986). It is thought that gestures lighten the load on the working memory, even if not referring to the physical environment. Speakers create a mental model of a real object which the gesture helps make present by simulating some of the salient physical features of the object (Ping & Goldin-Meadow, 2010), it is a process akin to the creation of metaphors. According to Hostetter, Alibali and Kita (2007) in conceptualizing, the mental image the speaker holds is broken down and reorganized into smaller chunks (of more salient features). By gesturing, these salient features are externalized, decreasing the working memory load and allowing for other information to be processed (Wagner,

Nusbaum & Goldin-Meadow, 2004). The effect is an overall reduction in working memory load. Thus, harder tasks will elicit more gestures, as a means of coping with the overall cognitive load.

It would seem that there are a number of reasons that might lead a speaker to gesture more, but all indicate more complexity, either as the idea takes shape or the speech is planned. If more demanding tasks, either because of conceptualization or formulation issues, lead speakers to gesture more, then gesture rates could be a good indicator of the difficulty of the task.

4. The study

This study investigated (i) the difficulty of three narration tasks, based on three different modalities of input: textual, aural and visual, as perceived by the participants and (ii) the adequacy of various parameters used to identify the difficulty differences between the three tasks. In addition to more traditional linguistic measures we sought to identify whether gesture-based parameters would be adequate to also evaluate task difficulty. The participants were ten university students from Hong Kong, in all cases English –the focus of this study– was their second language (L2). First the study identified the task modality that imposed the greatest difficulties on participants, as self-reported by the participants (an indication of the cognitive load imposed by the task). Then we focused on analysing the data using various traditional linguistic and gesture-based evaluation parameters to establish the match between the self-reflected difficulty rankings and those calculated. This allowed us to identify the most adequate parameters to give an indication of the difficulty of the tasks.

4.1. Participants

The data collection was based on convenience sampling, which was found to be a feasible option for a pilot study despite its limitations (Farrokhi and Mahmoudi-Hamidabad, 2012). Twelve students at the researcher's place of work answered calls to participate in this study, some of them already known to the researcher. Two participants were rejected as their mother tongue was Mandarin, not Cantonese. All were students at an English medium university in Hong Kong raised in a bilingual environment (Cantonese/English). All participants had attended an English medium secondary school and some had also attended an English medium primary school; their level of English was C1 or above –on the proficiency scale of the Common European Framework of

Reference for Languages (Council of Europe, 2001)– a requirement of their courses. Sixty per cent were female and all were under 25 years old.

4.2. Procedure

The participants were given stories in different modalities, a text, an audio and a video, and were asked to narrate them. In all three cases the speaker had to read, hear or watch approximately three minutes of input and then recount the story with as much detail as possible. In all cases the narration took place immediately after each type of input was given. Following the narrations, participants were asked to reflect on the tasks and report which had been the most difficult and why. Participants were allowed to read, listen to or watch the input as many times as necessary but there was no preparation time. Similarly, they were allowed to narrate the stories for as long as necessary. This was done to minimise variables relating to communicative stress, (Skehan, 2001) and resource depleting factors (Robinson, 2011).

In order to relate this study to existing gesture studies, for the audio and video the *Tweety and Sylvester* stories (a television cartoon) were used. In these stories Sylvester (a cat) is trying to capture and eat Tweety, a bird living with his owner (Granny). Based on the gestures work pioneered in McNeill's lab (http://mcneilllab.uchicago.edu/), we selected the first half of the *Canary Row* episode (Freleng, 1950) for the muted video input. The aural story was based on a similar *Tweety and Sylvester* episode scripted by the research team. The written texts were versions of the fable of *The Lion and the Mouse* by Aesop (n.d.) downloaded from the internet. All stories had been designed to be of a similar length and to contain a similar number of words, characters and events.

The procedure and general objective of the research was explained to participants and their consent obtained, although we did not specifically mention we were investigating measures of cognitive load such as gestures or recall capabilities. Each session was video-recorded and the speech and gestures in the recordings were transcribed. *PRAAT* (a speech transcription software) was used to note down the speech, which was added to *ELAN*, a multimodal transcription software allowing for frame by frame analysis of gestures, where the gesture transcription was carried out.

4.3. Analysis

Each participant narrated three stories in English, therefore a total 30 narrations of various lengths were transcribed and analysed, and the comments from the 10 interviews were recorded

and noted down (but not transcribed). The data from *ELAN* was then analysed using *WordSmith* (a Corpus analysis software) and *Excel and JASP* (statistical tools). From the speech transcripts, we manually calculated the number of events recalled, clauses and interruptions, other elements were calculated with *WordSmith*. From the gesture transcripts we manually calculated the total number of gestures produced. Manual calculations were repeated by a second researcher and disagreements were discussed until all issues were resolved.

A number of measures were chosen to evaluate accuracy, fluency and complexity, as indicators of task complexity (Robinson, 2011; Skehan, 2009). The selection was based on results from previous studies (Khawaja, Chen & Marcus, 2010; Gilabert, Barón, & Levkina, 2011). The full list of parameters is given in Table 1 and explained below.

Table 1. - List of evaluating parameters

Features Measured	**Parameter**	**Manual / computed**	**Task complexity element measured**
Recall	% events narrated	Manual	Accuracy
Speech rate	clauses/minute	Manual/computed	Fluency
Speech rate	tokens/minute	Computed/computed	Fluency
Lexical density	tokens/clause	Computed/manual	Complexity
Lexical density	type/token	Computed/computed	Complexity
Token complexity	word mean length	Computed	Complexity
Disfluency	eh/token	Manual/computed	Fluency
Disfluency	interruptions/token	Manual/computed	Fluency
Gesture rate	gesture/token	Manual/computed	Complexity?
Gesture rate	gesture/clause	Manual/manual	Complexity?

For each feature measured we aimed to find two evaluating parameters (not always possible), these are described below:

Percentage of events narrated: each story was divided into salient events. The number of events mentioned was noted and calculated as a proportion of the total number of events.

Number of clauses per minute: the transcripts of the narrations were divided into clauses, following the definition provided by Robinson (2011). Repairs were not double counted. The number of clauses was divided by the length of each narration.

Number of tokens per minute: number of words divided by the length of each narration, including interrupted words, and fillers such as ‘eh/ehm’, ‘you know’, etc.

Number of type of tokens per total number of tokens (token type ratio): the number of different types of words, including interrupted words, and fillers such as ‘eh/ehm’, ‘you know’, etc. divided by the total number of words in the narration.

Word mean length: the average number of letters in each word.

Number of ‘eh/ehm’ fillers per number of tokens: only fillers ‘eh/ehm’ (‘uh/uhm’, ‘ah/ahm’ and similar) were taken into account for this count. Other expressions such as ‘you know’, ‘I guess’ were not included in this study.

Number of interruptions per token: number of mid-word and end-of-word interruptions (signalled by a closing of the glottis) divided by the total number of tokens.

Number of gestures per token and per clause: arm and hand movements co-occurring with speech divided by either the number of words or the number of clauses. We excluded adaptors, hand and arm movements that answer a physical need.

It was expected that the easiest the task is for the participant, the better their speech would be, in terms of complexity, fluency and correction. Therefore, the three tasks were ranked as either 1 = easiest, 2 = in the middle, 3 = hardest, according to the results from each measurement taken. For example, speed of speech, ‘number of tokens per minute’, is assumed to be higher with easier tasks. Therefore, for the results in the example given in Table 2, the easiest task seems to have been the text based one and the hardest the aural based one. The percentage difference between the highest and lowest values, as a percentage of the lowest value was calculated, 16% in this example. We compared the self-ranking difficulty, as reported by participants, with the results from our calculations. In this example, our calculated results match the self-ranking. For each measuring parameter we noted the number of matches, in this case three, and tested the correlation between self and calculated rankings.

Table 2 – Example.

Task	Tokens /minute	% difference (Highest-Lowest)/Lowest	Calculated ranking (1=easiest)	Self-ranking (1 = easiest)
Text	148.90		1	1
Aural	128.35		3	3
Video	144.47		2	2
		16%		

5. Results

During the interview, participants had been asked to order, by level of difficulty, the three tasks, all narrated in English. Six participants (60%) agreed that the text based one was the easiest task. The video task was the second easiest for most participants (60%) and the audio based task was the hardest for half of them (50%). However, three participants (33%) also ranked the text task as the hardest one and also three (33%) ranked the video as the easiest one (see Appendix 1).

We added up the numerical values of the difficulty ranking given to each task (1, 2 or 3) by the students and as analysed. Students self-ranked the audio as being the hardest task, with a total compounded value of 22 (out of 30, the maximum value if all ten participants had scored the same task as being the hardest with a 3) and the text the easiest (with a value of 17). When adding the ranking values for each task by parameter we found that seven out of the ten parameters also indicate the text was the easiest task. Just one parameter, 'word mean length', indicated it was the hardest. Five parameters suggested the audio was the hardest task and two parameters, 'type token ratio' and 'word mean length', identified it as the easiest. Six parameters indicated that the video task was in the middle in terms of difficulty. With the parameter 'tokens per clause' all three tasks showed the same degree of difficulty. The values for the parameters 'percentage of events related' and 'gestures per token' showed the largest differences between the tasks.

Appendix 1 shows the ranking of each task, according to the observations made and comparing these with the self-rankings given by each participant, matches are shaded in grey. As can be observed, out of the 30 measurements (three per participant for ten participants) only three of the parameters provide over 50% matches between the self and calculated ranking. These are 'eh per token' (17 matches, 57%), and 'gesture per clause' and 'percentage of events related' (both with

16 matches, 53%). ‘Clauses per minute’, ‘tokens per minute’ and ‘gesture per token’ all had 13 matches each (43%). All other parameters gave under 30% matches.

Out of the ten parameters calculated, the average number of matches for each participant was 3.8. For each task at least three of the values calculated agreed with the self-ranking. This was the case for all participants, except for Participant 9 where the matches were just one or two per task.

The significance of these matches was statistically tested by running a Spearman’s correlation between the calculated rankings of the parameters and the self-rankings. Only the ‘eh per token’ parameter showed a significant strong positive correlation with the self-ranking values (rho = 0.542, p = 0.002). The two gesture based parameters gave non-significant medium strength correlations. Other parameters were neither significant nor correlated (see Table 3).

We also tested the correlation between the various parameters, in particular we were interested in a possible correlation between disfluencies, fillers or interruptions, and gestures. A strong correlation would indicate that the gesture might be aiding speech production at disfluencies. There did not seem to be a correlation between ‘eh per token’ and ‘gesture per token’ and ‘gesture per clause’ (rho = 0.048 and 0.000, p = 0.803 and 1 respectively), although there is a non-significant medium correlation between ‘interruptions per token’ and ‘gesture per clause’ (rho = 0.350, p = 0.058).

Table 3 – Spearman correlation matrix.

		% events related	clauses/ minute	tokens/ minute	tokens/ clause	type/ token	word mean length	eh/ token	interruptions/ token	gesture/ token	gesture/ clause
Self-ranking	Spearman's rho	0.001	0.109	0.166	-0.248	-0.090	-0.126	0.542 **	-0.188	0.284	0.286
	p-value	0.997	0.567	0.382	0.186	0.635	0.507	0.002	0.319	0.129	0.125
eh/token	Spearman's rho							—	0.100	0.048	0.000
	p-value							—	0.599	0.803	1.000
interruptions/token	Spearman's rho								—	0.243	0.350
	p-value								—	0.196	0.058

* p < .05, ** p < .01, *** p < .001

A Pearson test to evaluate the correlation between gestures and the number of clauses and gestures and the number of tokens indicates that there is a strong positive correlation in both cases (see Table 4).

Table 4. Correlation matrix for 30 pairs of values in each calculation.

		No. tokens	No. clauses
Gestures	Pearsons' r	0.914***	0.848***
	p-value	<0.001	<0.001

When comparing the results for the three tasks for each participant the variations measured were also very high. Variations of over 100% between the calculations of the easiest and the hardest tasks were observed in four parameters (see Appendix 1). This suggests considerable complexity differences between the tasks, at least for some students, as measured by those parameters. Other possibilities are an error in the transcriptions or calculations, however unlikely as data and calculations were all double-checked or perhaps, and more likely, outliers. A larger sample would clearly show the outliers, which we suspect, but cannot confirm, and strengthen the validity of the statistical test.

6. Discussion

The text based task seemed to be the easiest for this group of participants. From the feedback given by students, a number of possible reasons might explain this. Within their education systems participants were all used to processing textual, rather than visual data. Participants confirmed having studied in schools where up to one fourth of class-time was spent reading texts. Another reason might have been that this was a known story to participants and so they found it easier to recall. Finally, the reason that most participants gave was that in the textual task they were also provided with the tools to communicate the story, the words, making their task much easier. To counteract this ease, it might be significant to note that the text based task was the first one that all participants undertook, meaning there might have been some anxiety or uncertainty issues, raising the affective filter.

The audio task was reported to be the hardest, probably because it is seldom that we need to process data on speech alone, without any contextual or other nonverbal signals and these

participants had not been drilled in audio tasks. Although some of the participants reported having difficulties with the accent (slight Irish accent) it did not seem to be a common issue. As with the textual modality, the aural one also provided the participants with the tools to create the narration, many using the same tokens and expressions that had appeared in the audio.

The video modality, played without sound or subtitles, had the added difficulty –aside from having to remember the story, common to all three tasks– of having to generate the words in order to narrate the events. In the case of the video tasks, participants need to remember, understand and translate the concepts into words. In the case of the text and audio based tasks there is no need to understand, as long as participants are able to remember the words and repeat them in the same order. Despite these difficulties, 60% of participants considered this task the second easiest.

From the results and analysis, it would seem that the parameters 'eh per token', 'gesture per clause' and 'gestures per token' are the most useful ones to estimate the levels of difficulty imposed by these tasks. Out of the ten parameters two were gesture based and they are also two of the best performing ones.

The potential correlation between interruptions and gestures might indicate that the gesture is being used to help with speech formulation, however the lack of such a correlation between fillers and gestures suggests otherwise. The strong correlation between gestures and tokens and clauses indicates there is a relationship gesture-speech, but it is not possible to conclude whether the link gesture-speech is related to the gesture helping to formulate the speech or whether together with the speech it externalises the concepts. Further research investigating the type of gesture is recommended to clarify this point.

Although the parameter 'eh per token' seems to be a good indicator of task difficulty, it is likely that our results would show an even stronger correlation should we had included other fillers such as 'you know', 'and', 'actually', etc. These were not taken into account due to the complexity of identifying them as fillers or as a necessary part of the content. We observed that the use of 'eh' and 'ehm' might be slightly different, 'eh' perhaps signalling less complex difficulties (as suggested by Fox-Tree and Clark, 2002). Gestures might also be used differently with these two types of fillers, an area worth further exploration.

The parameter 'percentage events narrated' showed a surprisingly low statistical correlation, despite there being a relatively high percentage of matches between the self and calculated difficulty rankings. A larger sample might clarify this issue.

Parameters relating to the number of tokens per clause, ratio of token type and word mean lengths might have been skewed by the input itself, both the textual and aural tasks gave words and expressions that were often repeated by the participants. It is very likely that the weak correlations observed might be related to these interference from the input. To evaluate their strength as evaluating parameters, the study could be repeated using just textual and aural inputs and measuring the percentage of tokens that participants repeat from these.

Neither 'clauses per minute' nor 'tokens per minute' seem to be good indicators of difficulty. Perhaps a more reliable indicator would be lexical density, not tested by this study, excluding fillers and functional words in measuring tokens.

'Interruptions per token' also showed not to be a very reliable indicator. Research points out to a difference between mid-word interruption and end-of-word interruption, and whether the interruption is filled with a filler –eh, ehm– or followed by a pause (Lopez-Ozieblo, 2017, Seyfedinipur, 2006). Interruptions might signify errors but they also indicate a re-planning of the message, not necessarily because it was wrong, but because the speaker thought it inadequate. A better parameter might be just repairs, rather than interruptions (as suggested by Gilabert, Barón, & Levkina, 2011).

7. Conclusion

This study sought to identify the validity of existing and new evaluating parameters in identifying second language task difficulty, and so to indicate how the cognitive loads might have varied in these. The two objectives of the study were (i) to assess how input modality affects the difficulty of language tasks and (ii) assess the validity of parameters used to evaluate task difficulty, including new ones based on gestures. The three tasks were narrations based on inputs that were either textual, aural or visual. Ten Cantonese speakers of English as their L2 participated in the study. Participants' rankings of the difficulty of each task were correlated to the calculated values from the parameters tested and the correlations further studied.

The three tasks presented slightly different speech processing issues by reason of the differences in modality of input. All tasks had a clear macrostructure, therefore the stress on the Conceptualizer should have been similar. All three tasks presented the same number of characters. The audio and text based tasks provided the words in the input, therefore also minimising the stress on the Formulator (Skehan, 2009). The video task, on the other hand, would have taxed the Formulator,

as no words were provided in the input. It might have also taxed the Conceptualizer in that the events described could be interchangeable, therefore an effort had to be made to remember the sequence of events. We expected more speech issues —disfluencies, slower speech, less complexity— in the video based narration, but also more gestures due to the connection between gesture and visual processing (Baddeley, 1986) as well as the increased load on the Conceptualizer (Wagner et al., 2004). These expectations were only confirmed in the results from participants who reported the video-based task as being the hardest. Overall, more gestures were observed in all the tasks that had been ranked as the hardest. This suggests that gestures are not just the result of visual input. From an educational point of view, these results indicate that teachers in this context, higher education in Hong Kong, should provide text based input to ease the learning process. On the other hand, industry requires processing of non-textual data on a daily basis and reporting based on aural input and visual assessments. Lack of training in these skills might provide an explanation for our results and so aural and visual input based tasks should also be part of the curriculum.

Our results suggest that most individual parameters tested might not be very reliable indicators of task difficulty levels. From a statistical point of view, only the evaluating parameter 'eh per token' provided values strongly correlated with the difficulty self-ranking values. However, when we analysed the other evaluating parameters, we observed that 'percentage events related' and 'gesture per clause' might also be adequate parameters. It is strongly recommended that the study is repeated with a larger sample to identify whether the correlations of these two other variables might also prove to be statistically significant.

It was expected that gestures could be an accurate measure of the difficulty of the data, as difficulty imposes a heavier cognitive load and gestures have been confirmed to ease that load. It is not clear from our study whether these gestures are generated together with the speech at the Conceptualizer (Lopez-Ozieblo & McNeill, 2017) or whether they activate the Formulator, Lexical Retrieval Hypothesis (Rauscher, Krauss & Chen, 1996). Rauscher, Krauss and Chen's hypothesis (1996) suggests that gestures facilitate access to the lexicon, therefore they would be aiding the Formulator or even the Articulator. In this study, our second language speakers are presumably not struggling to conceptualize the idea if the language has been understood (as the macro-structure of the stories were given to the participants). As the words were already given in the textual and aural tasks, there should be a reduction of complexity in these two tasks, compared to the video one, and if anything fewer gestures altogether, as neither conceptualizing nor formulating was

particularly taxed. However, participants might still have formulating and articulation issues as they are requested to use the L2 to narrate the stories.

If gestures are used to aid speech processing we could expect to see more gestures together with disfluencies. This did not seem to be the case, there is no correlation between 'eh per token' and either gesture parameter, although there is a potential correlation between 'interruptions per token' and 'gestures per clause'. It would seem that at least some of the gestures might be helping ease the cognitive load, rather than just priming speech, thus giving an indication of the complexity of the task.

As our inputs were in either word based or visual based modalities we cannot comment on Baddeley and Hitch's (1974) dual processing theories. In all tasks one of the processors, the one coding verbal data or the one coding and retrieving images, would have been at work. It is suggested that future studies combine input modalities and compare the use of gestures with single-modality-input narrations.

Aside from the small sample size, the number of cofounding variables that we did not control for is a key limiting factor in this study. In particular, the content of the stories, although we tried to design the stories to be as similar as possible in terms of number of characters, number of events and potential imagery created by them. After the debriefing sessions with the participants, it became obvious that they considered the video story as more complicated because the events were interchangeable, a factor we had missed. Another limitation is that fillers were counted as tokens. It was difficult not to do so, as gestures might also occur during these. Therefore, eliminating fillers would have meant eliminating the gestures co-occurring with them. Clearly, further research is needed to identify the relationship between the fillers and the gestures and whether these serve a lexical retrieval function or a conceptualizing one.

Despite all of the limitations, this pilot study has been useful to highlight the possibility of using gestures for second language performance evaluation. Although most of these ten parameters did not perform very well individually, together they did match the overall difficulty self-rankings given by participants. When measuring the difficulty of tasks based on multiple modalities, the best basket of evaluating parameters might be a combination of the stronger correlating parameters, together with those identified as strong matches manually. Our recommendation would be: 'eh per token', 'gesture per clause' and 'percentage events related'. Further testing of the adequacy of these parameters is underway by the researching team.

Acknowledgments

This study would not have been possible without the collaboration of students from the Hong Kong Polytechnic University and the work of Mr. Cyril Lim.

Thank you to the organizers of the *International Conference on ESP, New Technologies and Digital Learning* where parts of this paper were originally presented.

Funding: This study was funded by the Hong Kong Polytechnic University.

The PI of the project is a member of the Research Centre for Professional Communication in English of the Hong Kong Polytechnic University, the mission of which is to pursue applied research and consultancy so as to deepen the understanding of professional communication and to better serve the communicative needs of professional communities. This project is intended to fulfil in part this mission.

References

Aesop/Esopo (n.d). *The lion and the mouse.* Retrieved from: www.cpalms.org/Uploads/resources/.../TeachingLionandtheMouse.docx

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Baddeley, A. D. (1999). *Essentials of human memory*. Psychology Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation, 8*, 47-89.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123*(1-2), 1-30.

Candlin, C. (1987). Towards task-based language learning. *Language learning tasks*, 5-22.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*, 293–332.

Chen, C. M., & Wu, C. H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education, 80*, 108-121.

Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M.A., Taib, R., Yin, B. & Wang, Y. (2012). Multimodal behavior and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(4), 22.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speech. *Cognition, 84*, 73-111.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, U.K: Press Syndicate of the University of Cambridge.

Diao, Y., Chandler, P., & Sweller, J. (2007). The effect of written text on comprehension of spoken English as a foreign language. *The American journal of psychology*, 237-261.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

Engelkamp, J., & Cohen, R. L. (1991). Current issues in memory of action events. *Psychological Research, 53*(3), 175-182.

Engelkamp, J., & Zimmer, H. D. (1985). Motor programs and their relation to semantic memory. *German Journal of psychology*.

Farrokhi, F., & Mahmoudi-Hamidabad, A. (2012). Rethinking convenience sampling: Defining quality criteria. *Theory and practice in language studies, 2*(4), 784.

Ford, N., & Chen, S. Y. (2000). Individual differences, hypermedia navigation, and learning: an empirical study. *Journal of educational multimedia and hypermedia, 9*(4), 281-311.

Freleng, F. (director) (1950). *Canary row* [Animated Film]. New York: Time Warner.

Frick-Horbury, D., & Guttentag, R. E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *The American journal of psychology*, *111*(1), 43.

Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. *Second language task complexity: Researching the cognition hypothesis of language learning and performance*, 105-140.

Goldin-Meadow, S. (2003). Thought before language: Do we think ergative. Language in mind: *Advances in the study of language and thought*, 493-522.

Goldin-Meadow, S. (2010). When gesture does and does not promote learning. *Language and cognition, 2*(1), 1-19.

Hostetter, A. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*, 297-315.

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes, 22*(3), 313-336.

Khawaja, M. A., Chen, F., & Marcus, N. (2010). Using language complexity to measure cognitive load for adaptive interaction design. *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 333-336). ACM.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture: Window into thought and action* (pp. 162–185). New York, NY: Cambridge University Press.

Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes, 24*(5), 761-775.

Kollöffel, B. (2012). Exploring the relation between visualizer–verbalizer cognitive styles and performance with visual or verbal learning material. *Computers & Education, 58*(2), 697-706.

Kruger, J. L., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research, 7*(5).

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167-196.

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press.

Lopez-Ozieblo, R. (2017). Cut-offs and Gestures: Analytical Tools to Understand a Second Language Speaker. *Professional and Academic Discourse: an Interdisciplinary Perspective, 2*, 60-68.

Lopez-Ozieblo, R. & McNeill, D. (2017). Exchange on gesture-speech unity: What it is, where it came from. In R. Breckinridge Church, Martha W. Alibali and Spencer Kelly (Eds.), Why Gesture?: *How the hands function in speaking, thinking and communicating* (pp. 103-125). Philadelphia, PA: John Benjamins Publishing.

Mayer, R. E., & Massa, L. J. (2003). Three facets of visual and verbal learners: Cognitive ability, cognitive style, and learning preference. *Journal of educational psychology, 95*(4), 833.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52.

McNeill, D. (2015). *Why we gesture: The surprising role of hand movements in communication.* Cambridge: Cambridge University Press.

Miller, P. (2001). *Learning Styles: The Multimedia of the Mind.* Research Report.

Nesbit, J. C., & Hadwin, A. F. (2006). Methodological issues in educational psychology. *Handbook of educational psychology, 2*, 825-847.

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230-244.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

Paas, F., G., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71.

Ping, R., & Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science, 34*(4), 602-619.

Rauscher, F. B., Krauss R. M., & Chen Y. (1996). Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226–230.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual-task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition, 38*(4), 703-737.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*(1), 1-36.

Sasayama, S. (2016). Is a 'complex'task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal, 100*(1), 231-254.

Seyfeddinipur, M. (2006). *Disfluency: Interrupting speech and gesture.* Doctoral dissertation, Radboud University Nijmegen Nijmegen.

Skehan, P. (2001). Tasks and language performance. *Researching pedagogic tasks: Second language learning, teaching, and testing*, 167-185.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics, 30*(4), 510-532.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language teaching research, 1*(3), 185-211.

Surjono, H. D. (2015). The effects of multimedia and learning style on student achievement in online electronics course. *TOJET: The Turkish Online Journal of Educational Technology, 14*(1).

Toumpaniari, K., Loyens, S., Mavilidi, M. F., & Paas, F. (2015). Preschool children's foreign language vocabulary learning by embodying words through physical activity and gesturing. *Educational Psychology Review, 27*(3), 445-456.

Trofatter, C., Kontra, C., Beilock, S., & Goldin-Meadow, S. (2015). Gesturing has a larger impact on problem-solving than action, even when action is accompanied by words. *Language, cognition and neuroscience, 30*(3), 251-260.

Van Gog, T., Paas, F., Van Merrienboer, J. J. G., & Witte, P. (2005). Uncovering the problem-solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*, 237-244.

Wagner, S. M., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language, 50*(4), 395-407.

Appendix 1 - Results

Participant	Task	Self-ranking	% events related	clauses/ minute	tokens/ minute	tokens/ clause	type/ token	word mean length	eh/ token	interruptions/ token	gesture/ token	gesture/ clause	Matches *
1	text	1	2	1	1	1	2	3	1	1	2	2	5
F	aural	1	1	3	3	3	1	2	3	2	1	1	4
	video	1	3	2	2	2	3	1	2	3	3	3	1
2	text	1	1	1	1	3	2	3	2	1	2	2	4
F	aural	3	2	2	2	2	1	2	3	2	3	3	3
	video	2	3	3	3	1	3	1	1	3	1	1	0
3	text	3	1	3	1	1	2	3	3	3	2	3	5
M	aural	2	3	2	3	3	1	1	2	2	3	2	4
	video	1	2	1	2	2	3	2	1	1	1	1	5
4	text	1	1	3	2	1	3	2	2	2	1	1	4
M	aural	3	3	1	1	2	2	3	3	3	3	3	6
	video	2	2	2	3	3	1	1	1	1	2	2	4
5	text	1	1	1	1	3	2	2	1	2	1	1	6
F	aural	3	3	2	2	1	1	1	2	1	3	3	3
	video	2	2	3	3	2	3	3	3	3	2	2	4
6	text	3	1	1	1	3	2	2	3	1	1	1	2
M	aural	1	3	3	3	2	1	1	1	3	3	3	3
	video	2	2	2	2	1	3	1	2	2	2	2	7
7	text	1	1	1	1	3	2	3	1	2	1	1	6
M	aural	2	2	3	2	1	1	1	2	3	3	3	3
	video	3	1	2	3	2	3	2	3	1	2	2	3
8	text	2	2	2	2	1	2	3	2	1	1	1	5
F	aural	3	1	1	3	3	1	1	3	2	3	2	4
	video	1	3	2	1	2	3	2	1	3	2	3	2
9	text	3	1	2	1	1	2	3	2	1	2	2	1
F	aural	1	3	2	2	2	1	2	1	3	3	3	2

	video	2	2	1	3	3	3	1	3	2	1	1	2
10	text	1	1	1	1	3	1	3	2	2	1	1	6
F	aural	3	3	3	3	1	2	1	1	1	2	3	4
	video	2	2	2	2	2	3	2	3	3	1	2	6
Matches calculated-self			16	13	13	6	6	6	17	8	13	16	
% matches calculated-self			53%	43%	43%	20%	20%	20%	57%	27%	43%	53%	
% difference (highest-lowest /lowest value)**			85%	90%	40%	15%	59%	8%	179%	147%	122%	126%	
Added value of given ranking values (text)		17	**12**	**16**	**12**	20	20	*27*	**19**	**16**	**14**	**15**	
Added value of given ranking values (aural)		22	*24*	*22*	24	20	**12**	**15**	*21*	22	*27*	*26*	
Added value of given ranking values (video)		18	22	20	24	20	*28*	16	20	22	17	19	

Note: in grey the cases were the self-ranking matched the ranking according to the calculations.

* Number of matches between the calculated rankings and self-rankings

** These highest and lowest values refer to the actual measured values for each parameter (not given in this table, for further information on these please contact the author).